ANOVA using Python (with examples)

Renesh Bedre

What is ANOVA (ANalysis Of VAriance)?

- ANOVA test used to compare the means of more than 2 groups (t-test can be used to compare 2 groups)
- Groups mean differences inferred by analyzing variances
- ANOVA uses variance-based *F* test to check the group mean equality. Sometimes, ANOVA *F* test is also called omnibus test as it tests non-specific null hypothesis i.e. all group means are equal
- Main types: One-way (one factor) and two-way (two factors) ANOVA (factor is an independent variable)
- It is also called univariate ANOVA as there is only one dependent variable in the model. MANOVA is used when there are multiple dependent variables in the dataset.
- If you have repeated measurements for treatments or time on same subjects, you should use Repeated Measure ANOVA

Note: In ANOVA, group, factors, and independent variables are similar terms

ANOVA Hypotheses

- *Null hypothesis*: Groups means are equal (no variation in means of groups)
  H_0: $\mu_1=\mu_2=\ldots=\mu_p$
- *Alternative hypothesis*: At least, one group mean is different from other groups
  H_1: All μ are not equal

The null hypothesis is tested using the omnibus test (F test) for all groups, which is further followed by post-hoc test to see individual group differences.

ANOVA Assumptions

- Residuals (experimental error) are approximately normally distributed (Shapiro-Wilks test or histogram)

- homoscedasticity or Homogeneity of variances (variances are equal between treatment groups) (Levene's or Bartlett's Test)
- Observations are sampled independently from each other (no relation in observations between the groups and within the groups) i.e., each subject should have only one response
- The dependent variable should be continuous. If the dependent variable is ordinal or rank (e.g. Likert item data), it is more likely to violate the assumptions of normality and homogeneity of variances. If these assumptions are violated, you should consider the non-parametric tests (e.g. Mann-Whitney U test, Kruskal-Wallis test).

How ANOVA works?

- Check sample sizes: equal number of observation in each group
- Calculate Mean Square for each group (MS) (SS of group/level-1); level-1 is a degrees of freedom (df) for a group
- Calculate Mean Square error (MSE) (SS error/df of residuals)
- Calculate *F* value (MS of group/MSE)
- Calculate *p* value based on *F* value and degrees of freedom (df)

One-way (one factor) ANOVA with PythonPermalink

ANOVA effect model, table, and formula

The ANOVA table represents between- and within-group sources of variation, and their associated degree of freedoms, the sum of squares (SS), and mean squares (MS). The total variation is the sum of between- and within-group variances.
The *F* value is a ratio of between- and within-group mean squares (MS). *p* value is

estimated from *F* value and degree of freedoms.

$$y_{ik} = \mu + \alpha_i + \epsilon_{ik}$$

$$SS_T = SS_B + SS_E$$

Where, y_{ik} = k^{th} observation of i^{th} level of groups,
μ = overall population mean (unknown) ,
α_i = Main effect for groups (deviation from the μ) ,
ϵ_{ik} = Error,
i = levels for groups (i = 1,2...,p) ,
k = Observations or replicates for each group (k = 1,2...,r) ,

Source of variation	degree of freedom (Df)	Sum of squares (SS)	Mean square (MS)	*F* value	Significance
Group (between)	Df_B = p-1	SS_B	$MS_B = SS_B / Df_B$	MS_B / MS_E	*p* value
Residuals or error (within)	Df_E = p(r-1)	SS_E	$MS_E = SS_E / Df_E$		
Total	Df_T = pr-1	SS_T			

Where, $SS_B = \sum_i p_i(\bar{y_i}. - \bar{y}..)^2$,
$SS_E = \sum_{ik}(y_{ik} - \bar{y_i}.)^2$,
$SS_T = SS_B + SS_E = \sum_{ik}(y_{ik} - \bar{y}..)^2$,

ANOVA example

Example data for one-way ANOVA analysis tutorial, dataset

A	B	C	D
25	45	30	54
30	55	29	60
28	29	33	51

A	B	C	D
36	56	37	62
29	40	27	73

Here, there are four treatments (A, B, C, and D), which are groups for ANOVA analysis. Treatments are independent variable and termed as factor. As there are four types of treatments, treatment factor has four levels.

For this experimental design, there is only factor (treatments) or independent variable to evaluate, and therefore, one-way ANOVA method is suitable for analysis.

Note: If you have your own dataset, you should import it as pandas dataframe. Learn how to import data using pandas

```
import pandas as pd
# load data file
df = pd.read_csv("https://reneshbedre.github.io/assets/posts/anova/onewayanova.txt", sep="\t")
# reshape the d dataframe suitable for statsmodels package
df_melt = pd.melt(df.reset_index(), id_vars=['index'], value_vars=['A', 'B', 'C', 'D'])
# replace column names
df_melt.columns = ['index', 'treatments', 'value']

# generate a boxplot to see the data distribution by treatments. Using boxplot, we can
# easily detect the differences between different treatments
import matplotlib.pyplot as plt
import seaborn as sns
ax = sns.boxplot(x='treatments', y='value', data=df_melt, color='#99c2a2')
ax = sns.swarmplot(x="treatments", y="value", data=df_melt, color='#7d0013')
plt.show()
```

```
import scipy.stats as stats
# stats f_oneway functions takes the groups as input and returns ANOVA F and p value
fvalue, pvalue = stats.f_oneway(df['A'], df['B'], df['C'], df['D'])
print(fvalue, pvalue)
# 17.492810457516338 2.639241146210922e-05

# get ANOVA table as R like output
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Ordinary Least Squares (OLS) model
model = ols('value ~ C(treatments)', data=df_melt).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
# output (ANOVA F and p value)
               sum_sq    df         F    PR(>F)
C(treatments)  3010.95   3.0  17.49281  0.000026
Residual        918.00  16.0       NaN       NaN

# ANOVA table using bioinfokit v1.0.3 or later (it uses wrapper script for anova_lm)
from bioinfokit.analys import stat
res = stat()
res.anova_stat(df=df_melt, res_var='value', anova_model='value ~ C(treatments)')
res.anova_summary
# output (ANOVA F and p value)
                df   sum_sq   mean_sq         F    PR(>F)
C(treatments)  3.0  3010.95  1003.650  17.49281  0.000026
Residual      16.0   918.00    57.375       NaN       NaN

# note: if the data is balanced (equal sample size for each group), Type 1, 2, and 3 sums of squares
# (typ parameter) will produce similar results.
```

Interpretation

The *p* value obtained from ANOVA analysis is significant ($p < 0.05$), and therefore, we conclude that there are significant differences among treatments.

Note on *F* value: *F* value is inversely related to *p* value and higher *F* value (greater than *F* critical value) indicates a significant *p* value.

Note: If you have unbalanced (unequal sample size for each group) data, you can perform similar steps as described for one-way ANOVA with balanced design (equal sample size for each group).

From ANOVA analysis, we know that treatment differences are statistically significant, but ANOVA does not tell which treatments are significantly different from each other. To know the pairs of significant different treatments, we will perform multiple pairwise comparison (**post hoc comparison**) analysis for all unplanned comparison using **Tukey's honestly significantly differenced (HSD)** test.

Note: When the ANOVA is significant, post hoc tests are used to see differences between specific groups. post hoc tests control the family-wise error rate (inflated type I error rate) due to multiple comparisons. post hoc tests adjust the *p* values (Bonferroni correction) or critical value (Tukey's HSD test).

Tukey's HSD test accounts for multiple comparisons and corrects for family-wise error rate (FWER) (inflated type I error)

Tukey and Tukey-kramer formula

Alternatively, Scheffe's method is completely coherent with ANOVA and considered as more appropriate post hoc test for significant ANOVA for all unplanned comparisons. However, it is highly conservative than other post hoc tests.

```
# we will use bioinfokit (v1.0.3 or later) for performing tukey HSD test
# check documentation here https://github.com/reneshbedre/bioinfokit
from bioinfokit.analys import stat
# perform multiple pairwise comparison (Tukey's HSD)
# unequal sample size data, tukey_hsd uses Tukey-Kramer test
res = stat()
res.tukey_hsd(df=df_melt, res_var='value', xfac_var='treatments', anova_model='value ~ C(treatments)')
res.tukey_summary
# output
  group1 group2  Diff      Lower      Upper   q-value   p-value
0      A      B  15.4   1.692871  29.107129  4.546156  0.025070
1      A      C   1.6 -12.107129  15.307129  0.472328  0.900000
2      A      D  30.4  16.692871  44.107129  8.974231  0.001000
3      B      C  13.8   0.092871  27.507129  4.073828  0.048178
4      B      D  15.0   1.292871  28.707129  4.428074  0.029578
5      C      D  28.8  15.092871  42.507129  8.501903  0.001000

# Note: p-value 0.001 from tukey_hsd output should be interpreted as <=0.001
```

Above results from Tukey's HSD suggests that except A-C, all other pairwise comparisons for treatments rejects null hypothesis ($p < 0.05$) and indicates statistical significant differences.

Note: Tukey's HSD test is conservative and increases the critical value to control the experimentwise type I error rate (or FWER). If you have a large number of comparisons (say > 10 or 20) to make using Tukey's test, there may be chances that you may not get significant results for all or expected pairs. If you are interested in only specific or few comparisons and you won't find significant differences using Tukey's test, you may split the data for specific comparisons or use the *t*-test

Test ANOVA assumptions

- ANOVA assumptions can be checked using test statistics (e.g. Shapiro-Wilk, Bartlett's, Levene's test) and the visual approaches such as residual plots (e.g. QQ-plots) and histograms.
- The visual approaches perform better than statistical tests. For example, the Shapiro-Wilk test has low power for small sample size data and deviates significantly from normality for large sample sizes (say n > 50). For large sample sizes, you should consider to use QQ-plot for normality assumption.

Now, I will generate QQ-plot from standardized residuals (outliers can be easily detected from standardized residuals than normal residuals)

```
# QQ-plot
import statsmodels.api as sm
import matplotlib.pyplot as plt
# res.anova_std_residuals are standardized residuals obtained from ANOVA (check above)
sm.qqplot(res.anova_std_residuals, line='45')
plt.xlabel("Theoretical Quantiles")
plt.ylabel("Standardized Residuals")
plt.show()

# histogram
plt.hist(res.anova_model_out.resid, bins='auto', histtype='bar', ec='k')
plt.xlabel("Residuals")
plt.ylabel('Frequency')
plt.show()
```

As the standardized residuals lie around the 45-degree line, it suggests that the residuals are approximately normally distributed

In the histogram, the distribution looks approximately normal and suggests that residuals are approximately normally distributed